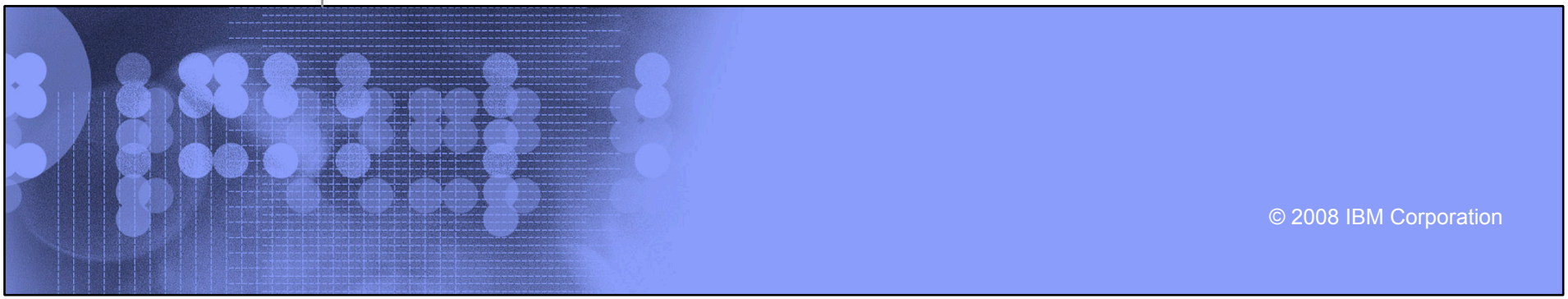
Compiler Optimization hints for maximizing performance on Blue Gene / P

Jerry Heyman, RTP
Technical Consultant
jheyman@us.ibm.com



Agenda

- Introduction
- -O3 v -O4
- -qhot and -qipa
- language unique options
- “fast” libraries

-O3 v -O4

- -O3 specifies the following as the default:
  - -qnostrict
  - loop invariant fp computations are moved outside of the loop

```
...
int i;
float a[100], b, c;
for (i=0; i < 100; i++) {
   if ( a[i] < a[i+1])
      a[i] = b + c;   // -O2 b+c is recalculated each time
                      //  -O3 b+c is moved outside the loop, done once
}
...
```

-O3 v -O4 (con't)

- fp computations may be rewritten
  - *a * b * c* could be rewritten *a * c * b*
- -qhot=noarraypad:level=0:nosimd:vector
  - high order loop analysis and transformation
  - -qhot=noarraypad:level=1:nosimd:vector (default for -O4)
- -qfloat=rsqrt
  - replaces
    divisi
    on by the result of a sqrt with multiplication by the reciprocal of the sqrt
- -qmaxmem=-1
  - use all the memory necessary to perform the optimization

-O3 v -O4 (con't)

- -O4 does everything in -O3 with the following additional options:
  - *sets -qarch=450d and -qtune=450*
    - *generates BG/P specific hw instructions rather than generic Power*
  - *sets -qcache values appropriate for BG/P hardware*
  - -qhot=noarraypad:level=1:nosimd:vector
    - *arraypad* allows compiler to increase size of array for better array-processing loops
    - level=1 performs default set of high order transformations
    - nosimd disables the conversion of loop array operations to calls to vector instructions
    - vector converts certain operations that are performed via loops (ex: sqrt or reciprocal sqrt) to a call into the Mathematical Acceleration Subsystem (MASS library)

-O3 v -O4 (con't)

- *sets the default for -qipa*

-qipa=inline=auto:level=1:missing=unknown:partition=medium:threads=auto

- *IPA = interprocedual analysis class of optimizations*
  - *inline means that all functions < 8K in size can be inlined*
  - *level=1 enables inlining and limited alias analysis*
  - *missing=unknown limits the amount of interprocedural optimization on functions outside the source file being compiled.*
  - *partition=medium is the amount of memory used to do interprocedural analysis – the larger the partition, the more optimization, the longer the compilation.*
  - *threads=auto allows the compiler to pick the number of threads based on machine load. threads=noauto does one thread per machine processor*

Additional -qhot options

- simd (defaults to nosimd)
  - converts certain operations that are performed in a loop on successive elements of an array into a vector instruction
- vector
  - [illegible] might want to use *novector* if the change in precision is unacceptable.

Additional -qipa options

- inline
  - *limit*=XXXX sets the maximum relative size of the function
  - relative size is based on th e size of the function, the number of calls to the function, and so on.
- *level* specifies the amount of ipa optimization to occur (default 1)
  - *level=0* performs minimal analysis and optimization(-O3)
  - *l ev el=2* performs full interprocedural data flow and alias analysis (-O5)
- more a link time option than a co m pile time option, but you still must compile with either -O4 or -O5

Other common options

- **-qalias**=<value>
  - allows the user to specify if pointers are referenced by multiple variables. Can be used to turn off very aggressive optimizations.
- **-qlist -qsource**
  - **shows both source code and generated assembler**
- **-qxref[=full]**
  - **to see where variables are referenced and used.**
  - **=full will identify variables not used.**

Language specific options

- FORTRAN
  - **-qessl** option replaces intrinsic functions with the ESSL version, must link with -lesslbg or -lesslsmpbg
  - **-qfixed** specifies how many columns of text are allowed in a line.
    - xlf/xlf77 defaults to 72
    - all others default to 90
    - for portability, you might want -qfixed=132
- C/C++
  - **-qinfo=pro:gen**
    - -qnoinfo is the default for C
    - -qinfo=lan:trx is the default for C++
      - lan is language level effects
      - trx is rounding off of hexadecimal floating point constants

'fast' libraries

- Located in ***/bgsys/drivers/ppcfloor/comm/fast/bin***

script	mpich built with	comp used build app
fast/bin/mpicc	xl, no debug	xl
fast/bin/mpixlc	xl, no debug	xl (identical to mpicc)
comm/bin/mpicc	gcc	gcc
comm/bin/mpixlc	gcc	xl

- Caveats
- ***The 'fast' scripts use***
  - ***libraries that are built with assertions turned off***
  - ***MPICH debug turned off***

'fast' libraries (con't)

- Recommendations
  - *build* **and** *test* with original (comm/bin/mpi*) scripts
  - make sure you have successful runs of app before switching
- Using these shaves roughly a microse c ond off of most communications calls (which can be 25% improvement)